Appendix

Appendix A1.1 Study characteristics: Hancock, 2002 (randomized controlled trial)

Characteristic	Description
Study citation	Hancock, C. M. (2002). Accelerating reading trajectories: The effects of dynamic research-based instruction. *Dissertation Abstracts International*, *63*(06), 2139A. (UMI No. 3055690)
Participants	The study involved 94 second-grade students who attended a single school. Out of this group, 48 students received the intervention and 46 were in the comparison group. The students were randomly assigned into intervention and comparison groups using block randomization procedures. All students in the second-grade were administered several initial measures. Student scores were rank-ordered within each classroom, and then each student was matched with a similarly performing student. Students were then randomly assigned to the intervention group and the comparison group within matched pairs. No information was reported regarding student ethnicity or gender, but 11% of the students in this school qualified for free or reduced-price lunch. There was no attrition.
Setting	The study took place in one elementary school in the Kyrene school district in Tempe, Arizona.
Intervention	In additional to the regular curriculum (including reading instruction), the intervention group received 25 minutes of supplemental instruction using *Read Naturally* four times a week for 11 weeks. In each lesson, the first five minutes were spent on oral reading of a selected passage with a teaching assistant. The reading was timed for one minute and the total number of words read correctly was recorded on a graph. The last 20 minutes involved repeated oral reading of curriculum stories either individually or with a cassette tape. Once students practiced a passage eight times (three times with a cassette and five times individually), they did a timed reading with the teacher. If the student achieved mastery (100 words read correctly with three or fewer errors), the student moved onto another passage. Otherwise the cycle was repeated.
Comparison	In additional to their regular curriculum (including reading instruction), the comparison group students received supplemental instruction using *Connecting Math Concepts* curriculum (Level B). This program used worksheets, workbooks, coins, and games, and taught basic mathematics skills such as place value, money counting, time, addition, subtraction, and multiplication.
Primary outcomes and measurement	The author used the Peabody Picture Vocabulary Test (PPVT-III), the Word Use Fluency Test (WUF), and the Curriculum Based Measure: Cloze Probe and Test of Reading Fluency. The author used initial reading skills as a covariate to account for baseline differences between groups (see Appendices A2.1–2.2 for more detailed descriptions of outcome measures).
Teacher training	Six teaching assistants were trained over five days. Teaching assistants were observed modeling lessons during the training sessions and provided with written feedback. Teaching assistants were also observed once a week during the first phase, and at least once every three weeks during the second phase, receiving feedback as necessary.

Appendix A1.2 Study characteristics: Mesa, 2004 (quasi-experimental design)

Characteristic	Description
Study citation	Mesa, C. L. (2004). *Effect of Read Naturally Software on Reading Fluency and Comprehension.* Unpublished master's thesis, Piedmont College.
Participants	Twelve students from a single class were selected to participate because they had mastered certain decoding patterns. These students were matched into pairs based on their pre-intervention test scores (STAR Reading Test); one student was assigned to the intervention group and one to the comparison group.[1]
Setting	The study took place in one elementary school in Georgia.
Intervention	Students in the group left their regular class for *Read Naturally* (2001) computer instruction for 45 minutes, four days a week for three weeks. Students used the program independently unless they had a question or were attempting to pass a level, in which case they interacted with the teacher. The *Read Naturally* group worked with minimal teacher's supervision.
Comparison	The comparison group did not receive any special instruction and remained in the class with the regular classroom teacher.
Primary outcomes and measurement	The author administered the Oral Reading Fluency test. Two other outcomes, the STAR Reading Test and the Comprehension Reading Test were also used in the study, but have not been included in this review because sufficient information was not provided to evaluate face validity and reliability of these tests (see Appendices A2.1–2.2 for more detailed descriptions of the outcome measure).
Teacher training	No information on teacher training is provided.

1. The pretest equivalency of the two groups on the Oral Reading Fluency measure was verified by the WWC.

Appendix A2.1 Outcome measures in the fluency domain

Outcome measure	Description
Oral Reading Fluency	The test measures the number of words read per minute minus any errors. The passage was a 113-word passage (as cited in Mesa, 2004).
Curriculum Based Measurement: Test of Reading Fluency	Students were given passages from Level B of the Test of Reading Fluency, which are based on several published curricula and are designed to represent general grade-level reading material. The total number of words read correctly was recorded (as cited in Hancock, 2002).

Appendix A2.2 Outcome measures in the comprehension domain

Outcome measure	Description
Vocabulary	
Peabody Picture Vocabulary Test (PPVT) III	A standardized, receptive vocabulary test that asks students to choose which one of four pictures corresponds to a test word spoken aloud (as cited in Hancock, 2002).
Word Use Fluency	The Word Use Fluency test measured students' expressive language skills. The tester verbally presented words to the student, who was asked to use the words in a sentence. Words were presented one at a time, and the next word was presented once a response was given. The task lasted one minute, and the total correct number of responses was provided (as cited in Hancock, 2002).
Reading comprehension	
Curriculum Based Measurement: Cloze Probe	Students read passages of text and fill in key missing words from three choices (as cited in Hancock, 2002).

Appendix A3.1 Summary of study findings included in the rating for the fluency domain[1]

Outcome measure	Study sample	Sample size (students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Read Naturally* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[3] (*Read Naturally* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Hancock, 2002 (randomized controlled trial)[7]								
CBM: Test of Reading Fluency	Second grade	94	117.38 (30.73)	112.38 (30.52)	5.00	0.16	ns	+6
Average[8] for fluency domain (Hancock, 2002)						0.16	ns	+6
Mesa, 2004 (quasi-experimental design)[7]								
Oral Reading Fluency[9]	First grade	12	80.00 (20.66)	74.33 (25.56)	5.67	0.23	ns	+9
Average[8] for fluency domain (Mesa, 2004)						0.23	ns	+9
Domain average[8] for fluency						0.19	na	+8

ns = not statistically significant
na = not applicable

1. This appendix reports findings considered for the effectiveness rating and the average improvement index.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
4. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
5. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
6. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
7. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formula the WWC used to calculate statistical significance. In the case of Hancock (2002) and Mesa (2004), no corrections for clustering or multiple comparisons were needed.
8. The WWC-computed average effect sizes for each study and for the domain across studies are simple averages rounded to two decimal places. The average improvement indices are calculated from the average effect size.
9. The *Read Naturally* group mean equals the comparison group mean plus the mean difference. The computation of the mean difference took into account the pretest difference between the study groups.

Appendix A3.2 Summary of study findings included in the rating for the comprehension domain[1]

Outcome measure	Study sample	Sample size (students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Read Naturally* group[3]	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*Read Naturally* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Hancock, 2002 (randomized controlled trial)[8]								
Construct: Vocabulary development								
PPVT	Second grade	94	118.11 (16.14)	117.79 (17.50)	0.32	0.02	ns	+1
Word Use Fluency	Second grade	94	53.10 (12.07)	50.42 (12.20)	2.68	0.22	ns	+9
Construct: Reading comprehension								
CBM: Cloze Probe	Second grade	94	22.70 (8.66)	23.37 (7.18)	–0.67	–0.08	ns	–3
Domain average[9] for comprehension (Hancock, 2002)						0.05	na	+2

ns = not statistically significant
na = not applicable

1. This appendix reports findings considered for the effectiveness rating and the improvement index.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. Means are adjusted for pretest. The authors used initial reading skills as a covariant.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
5. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formula the WWC used to calculate statistical significance. In the case of Hancock (2002), a correction for multiple comparisons was needed.
9. The WWC-computed average effect sizes for each study and for the domain across studies are simple averages rounded to two decimal places. The average improvement indices are calculated from the average effect size. For a single study included in the comprehension domain, the study average is equal to domain average.

Appendix A4.1 *Read Naturally* rating for the fluency domain

The WWC rates an intervention's effects in a given outcome domain as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative.[1]

For the outcome domain of fluency, the WWC rated *Read Naturally* as having no discernible effects. It did not meet the criteria for other ratings (positive effects, potentially positive effects, mixed effects, potentially negative effects, and negative effects) because no studies showed statistically significant or substantively important effects.

Rating received

No discernible effects: No affirmative evidence of effects.

- Criterion 1: None of the studies shows a statistically significant or substantively important effect, either *positive* or *negative.*

  Met. No studies showed a statistically significant or substantively important effect, either positive or negative.

Other ratings considered

Positive effects: Strong evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *positive* effects, at least one of which met WWC evidence standards for a strong design.

  Not met. Only one study met the WWC evidence standards for a strong design, and that study did not show statistically significant positive effects.

AND

- Criterion 2: No studies showing statistically significant or substantively important *negative* effects.

  Met. No studies showed statistically significant or substantively important negative effects.

Potentially positive effects: Evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect.

  Not met. No studies showed a statistically significant or substantively important positive effect.

AND

- Criterion 2: No studies showing a statistically significant or substantively important *negative* effect and fewer or the same number of studies showing *indeterminate* effects than showing statistically significant or substantively important *positive* effects.

  Not met. No studies showed a statistically significant or substantively important negative effect, but one study showed indeterminate effects.

(continued)

Appendix A4.1 *Read Naturally* rating for the fluency domain *(continued)*

Mixed effects: Evidence of inconsistent effects as demonstrated through either of the following criteria.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect, and at least one study showing a statistically significant or substantively important *negative* effect, but no more such studies than the number showing a statistically significant or substantively important *positive* effect.

  Not met. No studies showed a statistically significant or substantively important effect, either positive or negative.

OR

- Criterion 2: At least one study showing a statistically significant or substantively important effect, and more studies showing an *indeterminate* effect than showing a statistically significant or substantively important effect.

  Not met. No studies showed a statistically significant or substantively important effect, while one study showed indeterminate effects.

Potentially negative effects: Evidence of a negative effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *negative* effect.

  Not met. No studies showed a statistically significant or substantively important negative effect.

AND

- Criterion 2: No studies showing a statistically significant or substantively important *positive* effect, or more studies showing statistically significant or substantively important *negative* effects than showing statistically significant or substantively important *positive* effects.

  Met. No studies showed a statistically significant or substantively important positive effect.

Negative effects: Strong evidence of a negative effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *negative* effects, at least one of which met WWC evidence standards for a strong design.

  Not met. No studies showed a statistically significant or substantively important negative effect.

AND

- Criterion 2: No studies showing statistically significant or substantively important *positive* effects.

  Met. No studies showed statistically significant or substantively important positive effects.

1. For rating purposes, the WWC considers the statistical significance of individual outcomes and the domain-level effect. The WWC also considers the size of the domain-level effect for ratings of potentially positive or potentially negative effects. See the WWC Intervention Rating Scheme for a complete description.

Appendix A4.2 *Read Naturally* rating for the comprehension domain

The WWC rates an intervention's effects in a given outcome domain as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative.[1]

For the outcome domain of comprehension, the WWC rated *Read Naturally* as having no discernible effects. It did not meet the criteria for other ratings (positive effects, potentially positive effects, mixed effects, potentially negative effects, and negative effects) because no studies showed statistically significant or substantively important effects.

Rating received

No discernible effects: No affirmative evidence of effects.

- Criterion 1: None of the studies shows a statistically significant or substantively important effect, either *positive* or *negative.*

  Met. No studies showed a statistically significant or substantively important effect, either positive or negative.

Other ratings considered

Positive effects: Strong evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *positive* effects, at least one of which met WWC evidence standards for a strong design.

  Not met. Only one study met the WWC evidence standards for a strong design, and that study did not show statistically significant positive effects.

AND

- Criterion 2: No studies showing statistically significant or substantively important *negative* effects.

  Met. No studies showed statistically significant or substantively important negative effects.

Potentially positive effects: Evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect.

  Not met. No studies showed a statistically significant or substantively important positive effect.

AND

- Criterion 2: No studies showing a statistically significant or substantively important *negative* effect and fewer or the same number of studies showing *indeterminate* effects than showing statistically significant or substantively important *positive* effects.

  Not met. No studies showed a statistically significant or substantively important negative effect, but one study showed indeterminate effects.

(continued)

Appendix A4.2 *Read Naturally* rating for the comprehension domain *(continued)*

Mixed effects: Evidence of inconsistent effects as demonstrated through either of the following criteria.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect, and at least one study showing a statistically significant or substantively important *negative* effect, but no more such studies than the number showing a statistically significant or substantively important *positive* effect.

  Not met. No studies showed a statistically significant or substantively important effect, either positive or negative.

OR

- Criterion 2: At least one study showing a statistically significant or substantively important effect, and more studies showing an *indeterminate* effect than showing a statistically significant or substantively important effect.

  Not met. No studies showed a statistically significant or substantively important effect, while one study showed indeterminate effects.

Potentially negative effects: Evidence of a negative effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *negative* effect.

  Not met. No studies showed a statistically significant or substantively important negative effect.

AND

- Criterion 2: No studies showing a statistically significant or substantively important *positive* effect, or more studies showing statistically significant or substantively important *negative* effects than showing statistically significant or substantively important *positive* effects.

  Met. No studies showed a statistically significant or substantively important positive effect.

Negative effects: Strong evidence of a negative effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *negative* effects, at least one of which met WWC evidence standards for a strong design.

  Not met. No studies showed a statistically significant or substantively important negative effect.

AND

- Criterion 2: No studies showing statistically significant or substantively important *positive* effects.

  Met. No studies showed statistically significant or substantively important positive effects.

1. For rating purposes, the WWC considers the statistical significance of individual outcomes and the domain-level effect. The WWC also considers the size of the domain-level effect for ratings of potentially positive or potentially negative effects. See the WWC Intervention Rating Scheme for a complete description.

Appendix A5 Extent of evidence by domain

Outcome domain	Number of studies	Sample size		Extent of evidence[1]
		Schools	Students	
Alphabetics	0	0	0	na
Fluency	2	2	106	Small
Comprehension	1	1	94	Small
General reading achievement	0	0	0	na

na = not applicable/not studied

1. A rating of "moderate to large" requires at least two studies and two schools across studies in one domain, and a total sample size across studies of at least 350 students or 14 classrooms. Otherwise, the rating is "small."